Identification of genomic regions carrying a causal mutation in unordered genomes

Pilar Corredor Moreno, Ghanasyam Rallapalli, Carlos A. Lugo, Dan MacLean
The Sainsbury Laboratory, Norwich, UK

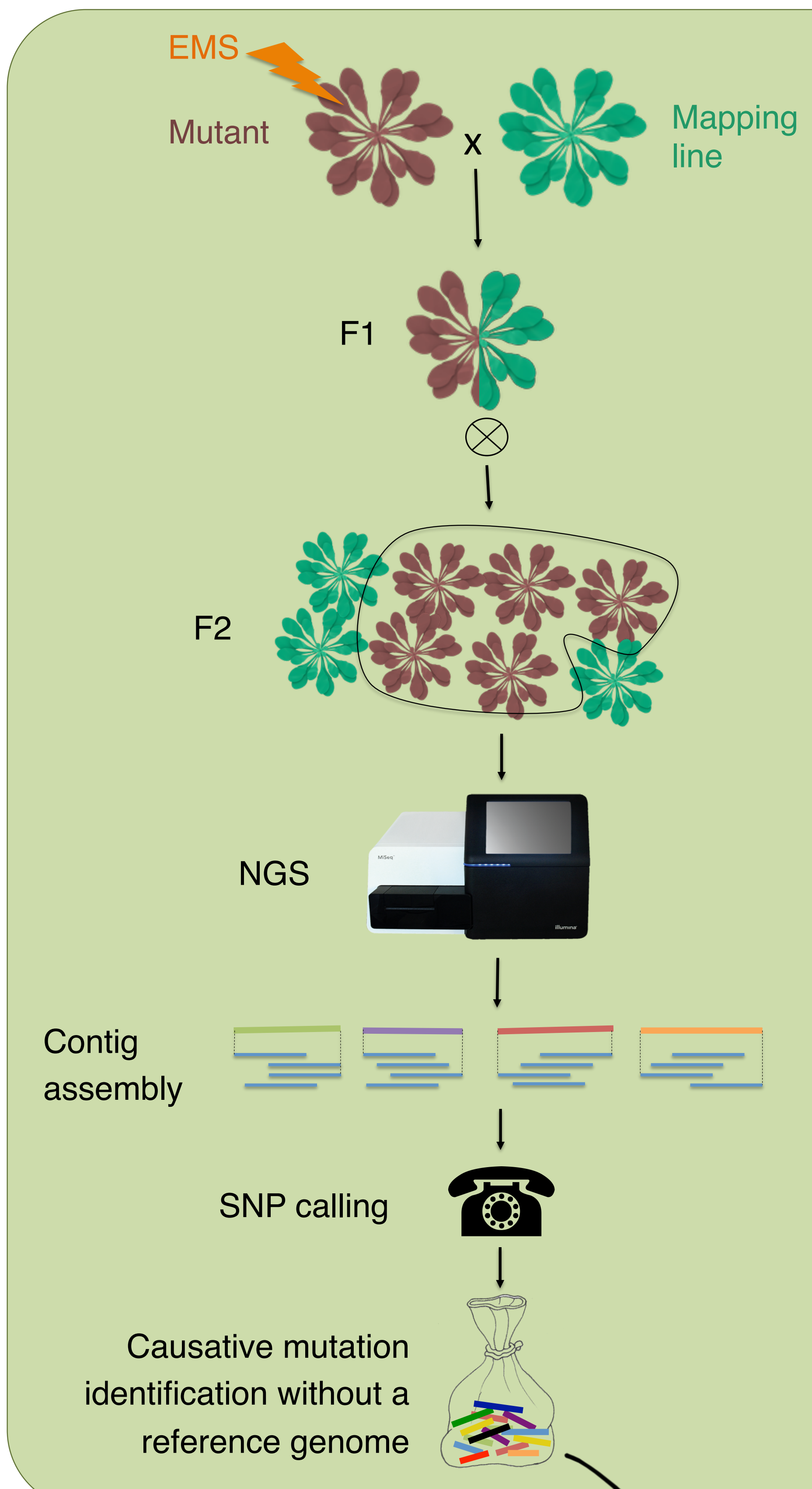
I'm a predoctoral intern at The Sainsbury Laboratory doing Bioinformatics in Team MacLean.

/pilarcormo/
SNP_Distribution_method

@PilarCorMo

Pilar.Moreno@tsl.ac.uk

Motivation

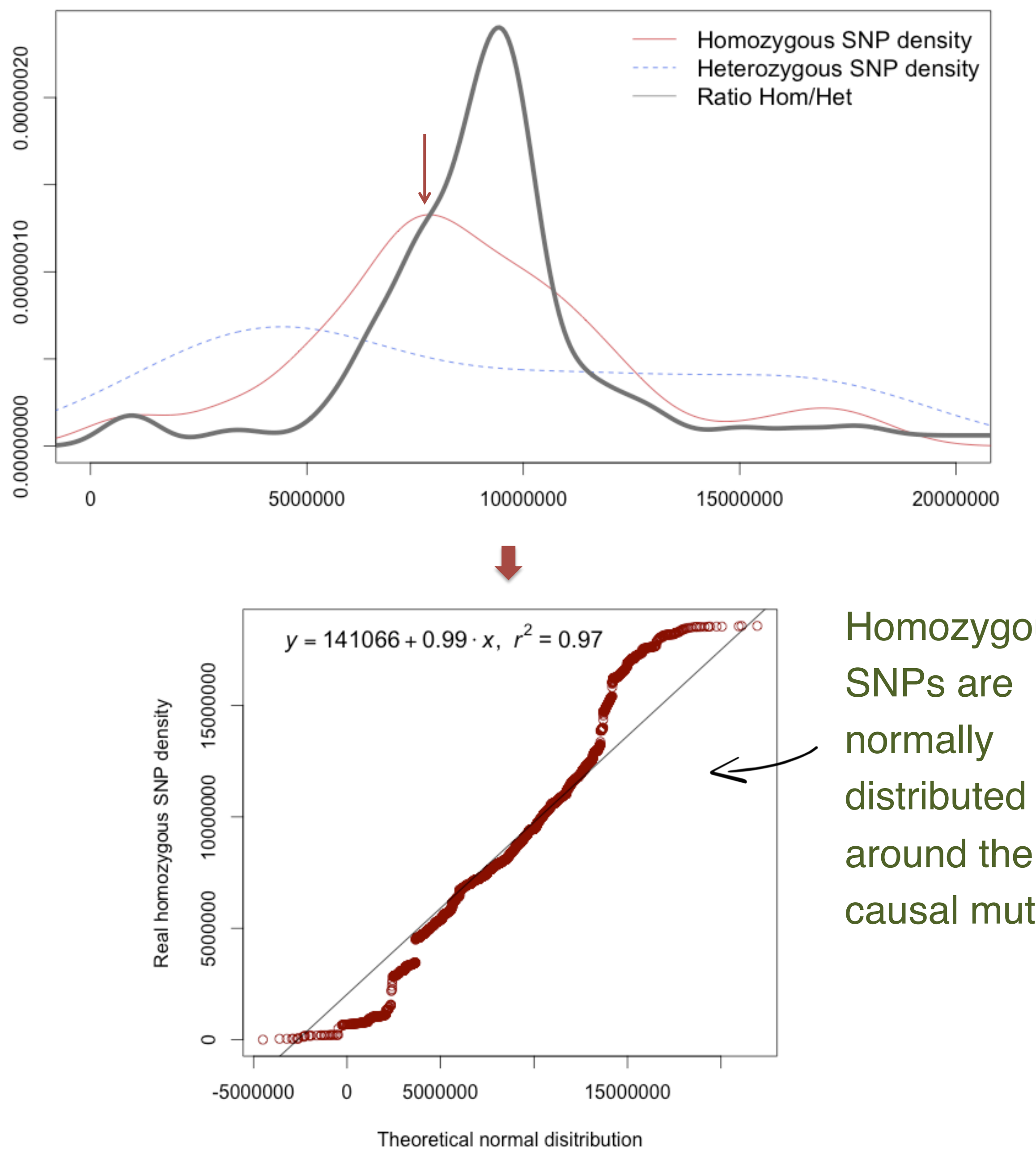

EMS
Mutant
x
Mapping line
F1
⊗
F2
NGS
Contig assembly
SNP calling
Causative mutation identification without a reference genome

Forward genetic screens have been a fundamental strategy to find genes involved in biological pathways in model species. Mutagenized individuals with a phenotype of interest are isolated and a recombinant mapping population is created by back-crossing to the parental line or out-crossing to a polymorphic ecotype.

The recombination frequency between the causal mutation and nearby genetic markers is low, so the alleles of these genetic markers will co-segregate with the phenotype-altering mutation while the remaining unlinked makers segregate randomly in the genome. Hence, allele distribution analysis can uncover these **low recombinant regions** to identify the location of the causal mutation.

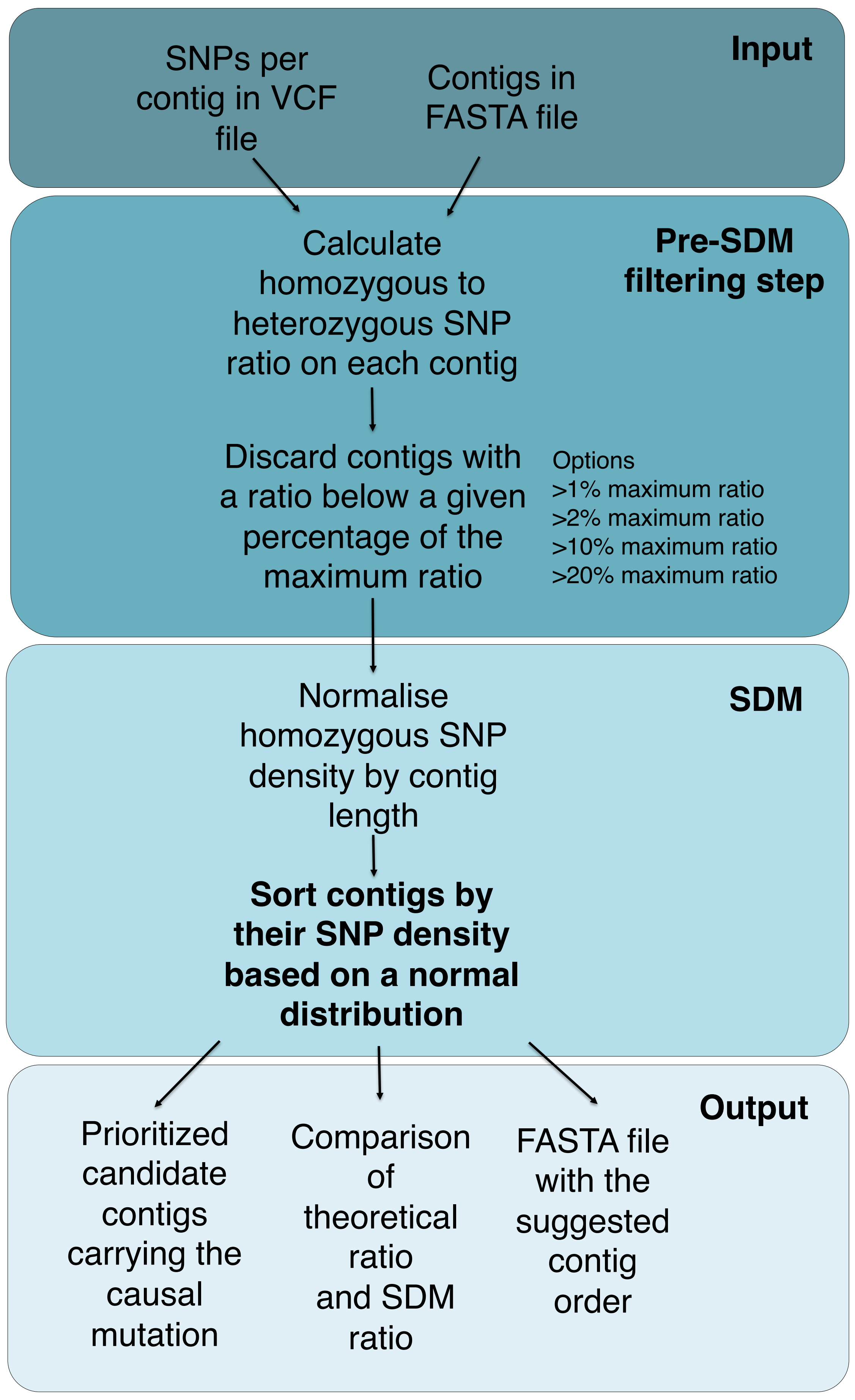
Traditional genetic mapping is a work intensive and time-consuming process but recent advances in high-throughput sequencing (HTS) have greatly accelerated the identification of mutations underlying mutant phenotypes in forward genetic screens. In the last few years, researchers have developed user-friendly tools for mapping-by-sequencing, yet they are **not applicable to organisms with non-sequenced genomes.**

SNP density plots revealed the homozygous SNP linkage around the causative mutation causing a high homozygous to heterozygous ratio signal where the mutation is located



Homozygous SNPs are normally distributed around the causal mutation!

SNP Distribution Method



Input

SNPs per contig in VCF file

Contigs in FASTA file

Pre-SDM filtering step

Calculate homozygous to heterozygous SNP ratio on each contig

Discard contigs with a ratio below a given percentage of the maximum ratio

Options
>1% maximum ratio
>2% maximum ratio
>10% maximum ratio
>20% maximum ratio

SDM

Normalise homozygous SNP density by contig length

Sort contigs by their SNP density based on a normal distribution

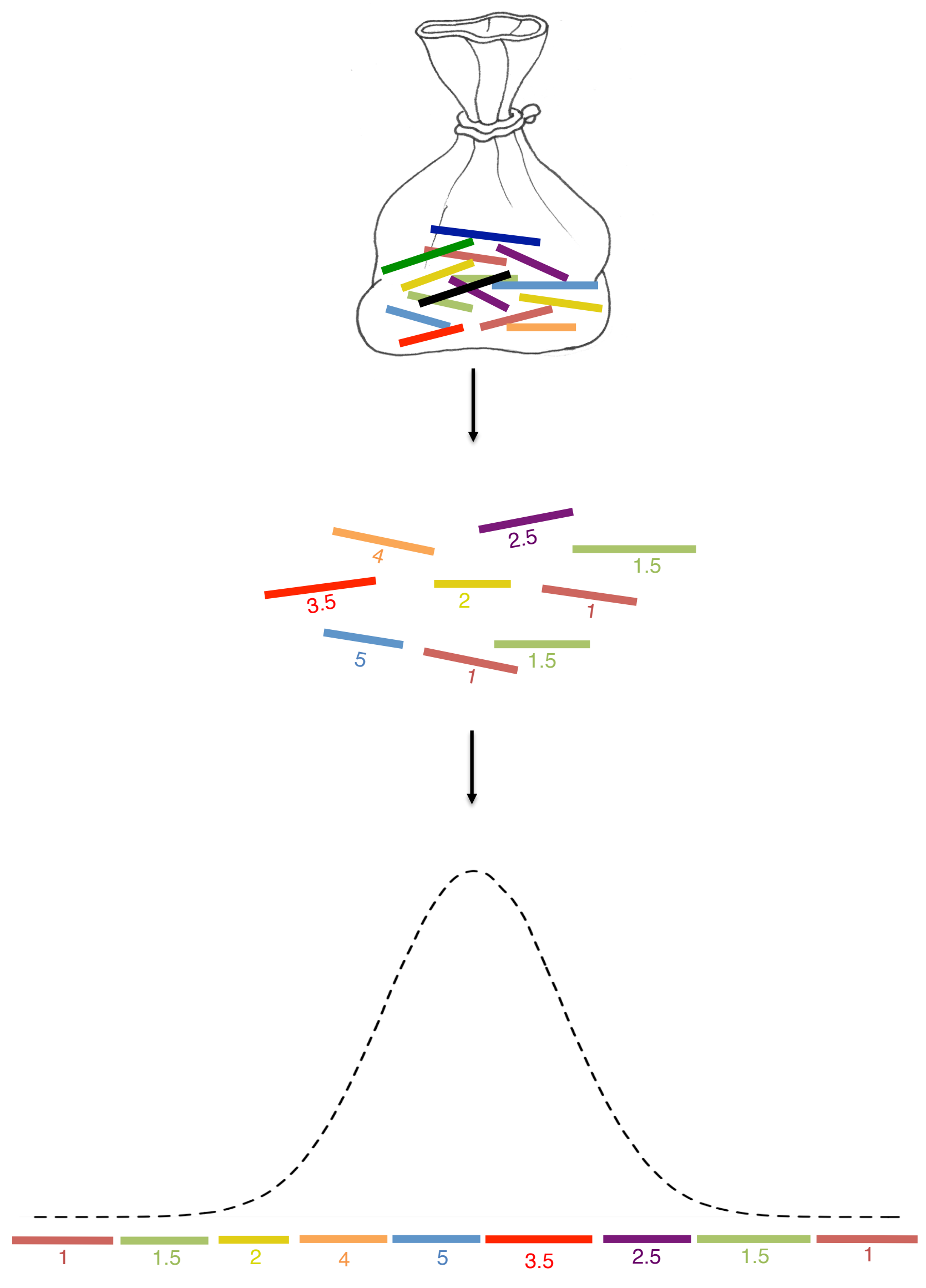
Output

Prioritized candidate contigs carrying the causal mutation

Comparison of theoretical ratio and SDM ratio

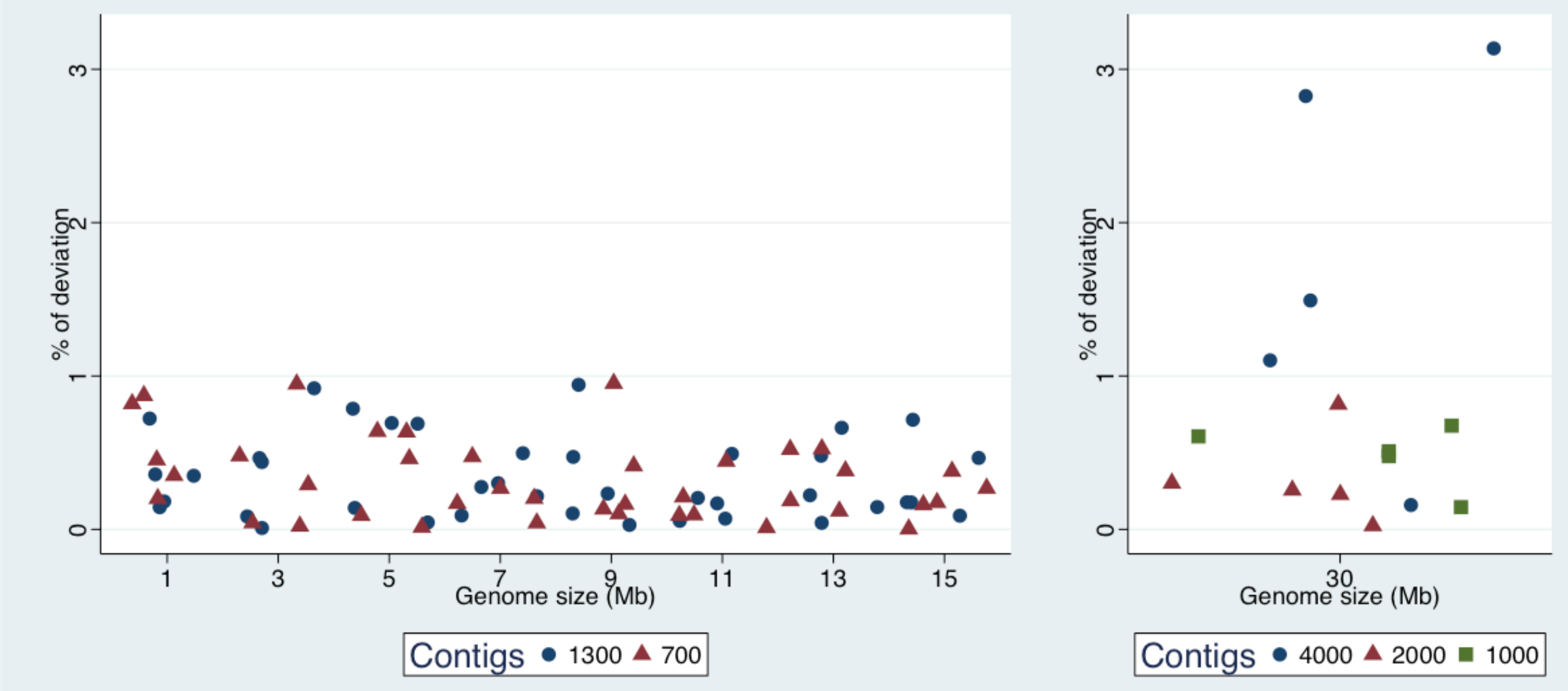
FASTA file with the suggested contig order

The SNP Distribution Method (SDM) is an approach for fast causative mutant identification based on a **simple reference-free contig assembly**. Instead of relying on a genome comparison, it focuses on the **SNP linkage around a causal mutation** and analyses the SNP distribution to identify the chromosome area where the putative mutated gene is located.



Modelling

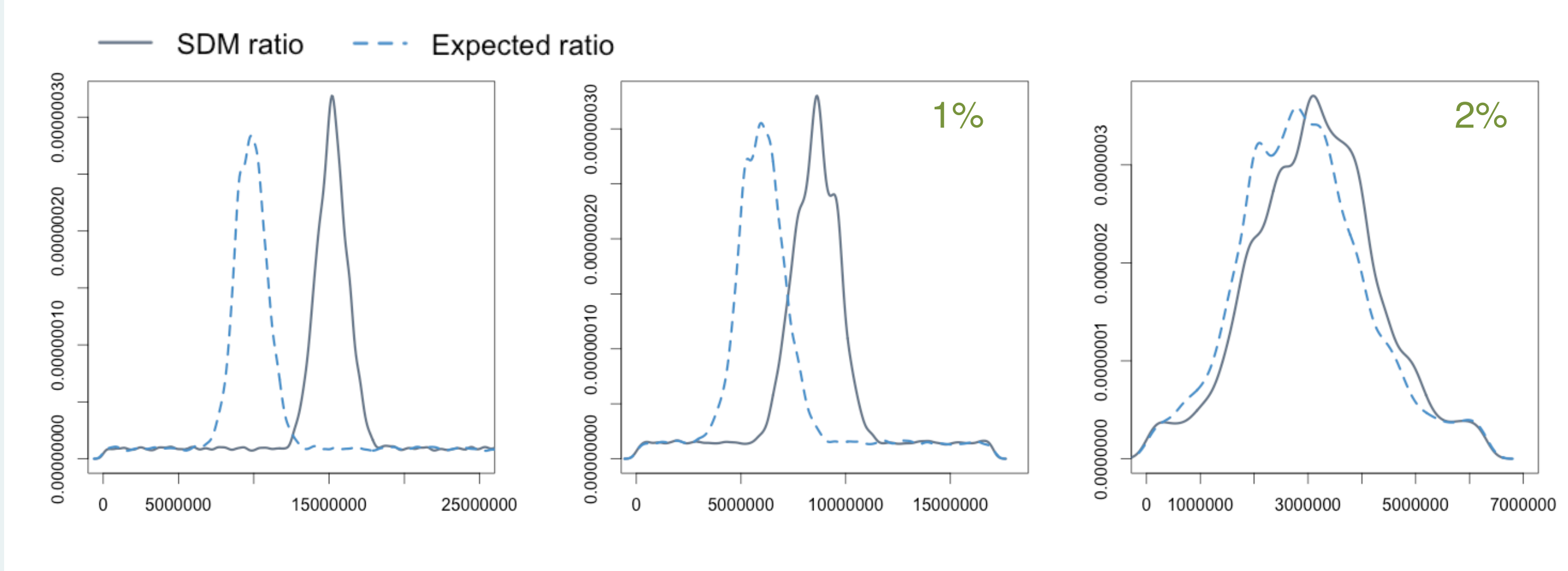
Model genomes are useful to help us develope our method and identifying its limitations. By using **idealised SNP distributions**, we can predict where a mutation is going to be located and estimate the deviation of SDM from this expected position. A normal distribution was used for the homozygous SNPs while heterozygous SNPs followed a uniform distribution. We created different model genomes based on *Arabidopsis thaliana* chromosome 1. We tested the effect of genome length and contig size on SDM performance.



We define the homozygous to heterozygous SNP ratio on contig n as:

$$Ratio_n = \frac{(\sum Hom) + 1}{(\sum Het) + 1}$$

Contigs that are located further away from the causal mutation have a constant homozygous SNP density due to recombination. The low ratio in these regions is used as a filter to focus on the genomic region where the mutation is likely to be found. **Contigs with a ratio falling below a given percentage of the maximum ratio will be discarded**.



Take home

- ✓ **Forward genetic screens** are very useful to identify genes responsible for particular phenotypes.
- ✓ Homozygous SNPs are **normally distributed** in the mutant genome of back-crossed and out-crossed individuals. We defined a theoretical SNP distribution used by SDM to identify the genomic region **where the causative mutation is located.**
- ✓ SDM does not rely on previously known genetic markers and can be used on extremely **fragmented genome assemblies**, even down to the level of long reads.